UG Programme: B. Tech (Honours) Computer Science and Engineering

(Data Science)

Course: DATA PREPROCESSING AND FEATURE ENGINEERING
Regulation: 2021 CBCS

Course Code: **Semester: IV**
Credits: 02 **L: T: P: 2:0:0**
CIE: UE: 50:50 **Maximum Marks: 100**
Contact Hours: 30

Pre-requisites: Python Programming/R Programming
Course Objectives:

The objective of the course is to
1. Understand the concept of feature engineering and different types of data.
2. Apply the suitable data pre-processing techniques on the given data for further analysis.
3. Understand the different feature selection techniques with suitable examples.
4. Enable the students to learn dimensionality reduction techniques with suitable examples.
5. Enable the students to learn text processing techniques with suitable examples.

Course Outcomes:

At the end of the course, students will be able to:

Course Outcomes	Description	Bloom's Taxonomy Level
CO1	Describe the different types of data and data handling technique.	Understanding (2)
CO2	Use different data preprocessing techniques on the given data.	Applying (3)
CO3	Use the different feature selection techniques on the given data.	Applying (3)
CO4	Illustrate the dimensionality reduction techniques with suitable examples	Applying (3)
CO5	Illustrate the text processing techniques with suitable examples.	Applying (3)
CO6	Compare the different data preprocessing, feature selection, dimensionality reduction techniques on the given data.	Analyzing (4)

Module I: Introduction to Feature Engineering **(6 Hours)**

Introduction to Feature Engineering, Why Feature Engineering, significance of Feature Engineering, Understanding the basics of data - Structured Data, Unstructured data, Quantitative versus Qualitative data, Encoding categorical variables : Encoding at the nominal level (One Hot Encoding, Dummy Encoding), Encoding at the ordinal level (Label Encoding, Frequency Encoding), Missing Data, what constitutes Missing Data, Different types of Missing Data, Different types of Missing Data handling techniques,.

Module II: Data Preprocessing **(6 Hours)**

Data Discretization (**Histogram analysis, Binning, Cluster analysis, Decision tree analysis, Correlation analysis),** Different types of data distribution, Data transformation: Normalization (Min-max normalization, Z score normalization, Normalization by decimal scaling), Data Standardization, Data reduction strategies: Sampling, Types of Sampling - Simple random sampling, Sampling without replacement, Sampling with replacement, Stratified sampling.

Module III: Feature Selection **(6 Hours)**

Feature selection, Types of feature selection, Statistical-based feature selection - Pearson correlation to select features, Feature selection using hypothesis testing - Interpreting the p-value, Ranking the p-value; Model-based feature selection - Using machine learning technique to select features, Feature selection by Tree-based model; Linear models and regularization - A brief introduction to regularization and different types of regularization.

Module IV: Dimensionality Reduction **(6 Hours)**

Introduction to Dimensionality Reduction techniques - Principal Component Analysis, Independent Component Analysis, Discrete Wavelet Transform, Linear Discriminant Analysis, Advantages of Dimensionality Reduction techniques.

Module V: Text Preprocessing **(6 Hours)**

Text data and Pre-processing - Lower Casing, Tokenization, Punctuation Mark Removal, Stop Word Removal, Stemming, Lemmatization, Part-of-Speech Tagging, Bag Of Words (BOW), Term Frequency - Inverse Document Frequency (Tf-IDF), Word2Vec.

Text Books:

1. Feature Engineering Made Easy by Sinan Ozdemir, Divya Susarla , January 2018, Published by O'REILLY.
2. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists, March 2018, Published by O'REILLY.

Reference Books and Links:

Feature Engineering and Selection: A Practical Approach for Predictive Models Max Kuhn and Kjell Johnson, 2019.

UG Programme: **B. Tech (Honours) Computer Science and Engineering (Data Science)**

Course: **DATA PREPROCESSING AND FEATURE ENGINEERING LAB**

Regulation: 2021 CBCS

Course Code: **Semester: IV**

Credits: 01 **L: T: P: 0:0:2**
CIE: UE: 70:30 **Maximum Marks: 100**
Contact Hours: 30

Pre-requisite:

Course Objectives:

1. Understand the concept of feature engineering and different types of data.
2. Apply the suitable data pre-processing techniques on the given data for further analysis.
3. Understand the different feature selection techniques with suitable examples.
4. To enable the students to learn dimensionality reduction techniques with suitable examples.
5. To enable the students to learn text processing techniques with suitable examples.

Course Outcomes:

At the end of the course, students will be able to:

Course Outcomes	Description	Bloom's Taxonomy Level
CO1	Describe the different types of data and data handling technique.	Understanding (2)
CO2	Use different data preprocessing techniques on the given data.	Applying (3)
CO3	Use the different feature selection techniques on the given data.	Applying (3)
CO4	Illustrate the dimensionality reduction techniques with suitable examples	Applying (3)
CO5	Illustrate the text processing techniques with suitable examples.	Applying (3)
CO6	Compare the different data preprocessing, feature selection, dimensionality reduction techniques on the given data.	Analyzing (4)

Expt. No. List of exteriments

1. Perform the following operations.
   a. Download the dataset from a public repository.
   b. Analyze the different types of variables in it.
2. Analyze the data for various variable characteristics such as,

a. Missing Data
b. Cardinality
c. Category Frequency
d. Distributions
e. Outliers
f. Magnitude.

3. Perform various imputation techniques for the missing data.

a. Mean-median imputation
b. Arbitrary value imputation
c. Frequent category imputation
d. Missing category imputation

4. Perform the variable encoding on the following type of data.

a. Categorical
b. Ordinal

5. Perform the process of Data discretization.

a. Histogram analysis
b. Binning, cluster analysis
c. Decision tree analysis
d. Correlation analysis

6. Normalize the data using different normalization techniques.

a. Min-max normalization
b. Z - score normalization
c. Normalization by decimal scaling

7. Select the optimal features from the dataset using various feature selection techniques.

a. Statistical-based feature selection
b. Pearson correlation to select features

8. Reduce the dimension of the dataset using Principal component analysis.

9. Perform dimensionality reduction using Linear discriminant analysis.

10. Download the dataset and perform the preprocessing steps as below

a. Lower Casing
b. Tokenization
c. Punctuation Mark Removal
d. Stop Word Removal
e. Stemming
f. Lemmatization.

11. Implement the Bag-of-words model.

12. Implement the following.

a. Term Frequency
b. Inverse Document Frequency (Tf-IDF)
c. Word2Vec model

Advanced Machine Learning

Subject Code : **Total Contact Hours : 45**

Credits : 03 **Hours per week : 03**

Course Objectives:

Implement Machine learning techniques using tensorflow
Assess ensemble models involved in machine learning concepts
Understand reinforcement learning concepts of machine learning
Test the built models using validation techniques
Deploy the machine learning models on cloud or local server

Unit-I: **(9 Hours)**

Advanced Machine learning with TensorFlow

Introduction, Tensorflow operations, declaring tensors, working with metrics, declaring operations, implementing activation functions, operations in computational graph, layering nested operations, working with multiple layers, implementing loss functions, implementing back propagation, working with batch and stochastic training, evaluating models, Implementing unit tests, multiple executors, productionalizing tensorflow.

Text Book – 4 - Chap 1,2, 10

Unit-II: **(9 Hours)**

Ensemble Methods

Bagging and Random forest, Bootstrap method, Bootstrap aggregation, Variable Importance, Boosting, AdaBoost, Boosting ensemble method, AdaBoost ensemble, CatBoost, Learning with ensembles, Implementing a simple majority vote classifier, Leveraging weak learners via adaptive boosting.

Text Book – 1 - Chap 5,

Text Book – 2 - Chap 7

Unit –III: **(9 Hours)**

Reinforcement Learning

Introduction, formal framework, different components to learn a policy, value based methods for RL, Q-learning, fitted Q-learning, Deep Q-networks, double DQN, dueling network architecture, distributional DQN, Multi step learning, concepts of generalization, feature selection, modifying objective function, hierarchical learning, bias-over fitting tradeoff.

Text Book – 3 - Chap 3, 4, 7

Unit IV: **(9 Hours)**

Model Evaluation and Hyper-Parameter Tuning

Streamlining workflows with pipelines, K-fold cross validation, Model performance measures, debugging algorithms with learning and validation curves, fine-tuning machine learning models via grid search, looking at different performance evaluation metrics, Ranking metrics, Classification metrics, regression metrics, Bootstrapping and Jackknife, Hold-out validation, difference between model validation and testing.

Text Book – 2 - Chap 6

Unit V: **(9 Hours)**

Machine Learning Deployment

Serializing fitted scikit – learn estimators, setting up a SQLite database for data storage, developing web application with Flask, turning the classifier into a web application, turning a regression problem into a web application, pickle model, deploying web application to a public server, Cloud deployment using AWS and Google.

Text Book – 2 - Chap 9

Course Outcomes:

On successful completion of the course, students will be able to:

Examine with advanced machine learning concepts.
Examine ensemble methods of machine learning.
Implement reinforcement learning in real world scenarios.
Demonstrate and deploy machine learning concepts
Deploy machine learning algorithms on cloud

Text Books:

1) Master Machine Learning Algorithms, Jason Brownlee

2) Deeper Insights into Machine Learning, Birmingham, Packt

3) An Introduction to Deep Reinforcement Learning, Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc

G. Bellemare and Joelle Pineau

4) Tensorflow machine learning cookbook, Nick McClure, Packt

Reference Books and links:

1) Advanced machine learning with python, John hearty, Packt

2) https://cloud.google.com/ml-engine/docs/deploying-models

3)https://towardsdatascience.com/simple-way-to-deploy-machine-learning-models-to-cloud-fd58b771fdcf

Advanced Machine Learning Lab

Subject Code: **Total Hours: 15**
Credits: 02 **L-T-P: 0-0-4**

List of Experiments:

1) Build a machine learning model for house price prediction analysis using lasso and ridge regression

2) Build a machine learning model on hand written digits and compare the models using evaluation techniques

3) Compare the differences between the accuracies obtained using ridge and lasso regression in first experiment

4) For the above build regression model, perform model evaluation, feature selection and parameter tuning

5) Build a classification model on heart disease UCI dataset using ensemble techniques

6) Compare the ensemble models built on heart disease data set and validate the same

7) Build a simple reinforcement learning model and use Montel Carlo learning to find the optimal combination of products using meal data with 4 ingredients and 9 products.

8) Build a Tic –Tac – Toe agent using Q-learning concept.

9) Financial Time Series Monte Carlo Simulation on S&P 500 stock data.

10) Deploy a regression model of first experiment using Flask and build a web api on the same.

11) Deploy the classification model of third experiment using amazon sage maker or as a pickle model as web api.

12) Deploy the classification model of third experiment using Google cloud or as a pickle model as web api.